Discrimination-Emphasized Mel-Frequency-Warping for Time-Varying Speaker Recognition

[1]Linlin Wang, [2]Thomas Fang Zheng, [3]Chenhao Zhang and [4]Gang Wang
Center for Speech and Language Technologies, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084
E-mail: { [1]wangll, [3]zhangchh, [4]wanggang }@cslt.riit.tsinghua.edu.cn, [2]fzheng@tsinghua.edu.cn
Tel/Fax: +86-10-62796393

***Abstract*— Performance degradation with time varying is a generally acknowledged phenomenon in speaker recognition and it is widely assumed that speaker models should be updated from time to time to maintain representativeness. However, it is costly, user-unfriendly, and sometimes, perhaps unrealistic, which hinders the technology from practical applications. From a pattern recognition point of view, the time-varying issue in speaker recognition requires such features that are speaker-specific, and as stable as possible across time-varying sessions. Therefore, after searching and analyzing the most stable parts of feature space, a Discrimination-emphasized Mel-frequency-warping method is proposed. In implementation, each frequency band is assigned with a discrimination score, which takes into account both speaker and session information, and Mel-frequency-warping is done in feature extraction to emphasize bands with higher scores. Experimental results show that in the time-varying voiceprint database, this method can not only improve speaker recognition performance with an EER reduction of 19.1%, but also alleviate performance degradation brought by time varying with a reduction of 8.9%.**

I. INTRODUCTION

Speaker recognition, also known as voiceprint recognition, is one kind of biometric authentication technology that can be used to automatically recognize a speaker's identity by using speaker-specific information contained in speech waves. Like all the other pattern recognition problems, it includes a training process (to obtain speaker models from utterances after feature-extraction) and a testing process (to determine the identity of a speaker-unknown utterance). This technology enables access control of various services by voice, including voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access of computers [1]. Apart from these commercial applications, it also has a prospect in forensic ones [2]. In all these typical situations, training and testing processes are usually separated by some period of time, which poses a possible threat to speaker recognition systems.

The time-varying issue has been mentioned many times since the birth of the word *voiceprint*. Although pioneer researchers believed identifiable uniqueness did exist in each voice just as that of fingerprints, they put forward this issue at the same time [3]. There was no evidence regarding the stability of speaker-specific information throughout time. In 1997, Sadaoki Furui summarized advances in automatic speaker recognition in decades and also left the way to deal with long-term variability in people's voice as an open question [1]. A similar idea was expressed in [4], where the authors argued that a big challenge to uniquely characterize a person's voice was that voice changes over time.

Performance degradation has also been observed in presence of time intervals in practical systems. F. Soong *et al*. [5] concluded from experiments that the longer the separation between training and testing recordings, the worse the performance. Kato and Shimizu [6] also reported a significant loss in accuracy between two sessions separated by 3 months and aging was considered to be the cause [7].

It is a generally acknowledged phenomenon that speaker recognition performance degrades with time varying. In spite of the fact that it is effective to update speaker models from time to time to maintain representativeness [1][5], few researchers have figured out reasons behind this phenomenon exactly.

From a pattern-recognition point of view, performance degradation results from mismatches between training and testing. All possible mismatches can be divided into two categories. One category is speaker-independent mismatches which originate from voice transmission outside speakers themselves. Environmental noise, echo, recording, and channel mismatches are of this category. The other category is mismatches in the speaking behavior of the same speaker (e.g., speaking style, speech content, time-related variability). Our research focuses on the time-related variability from the second category.

Due to absence of a proper longitudinal database, we created one that met this requirement with 16 recording sessions in a period of approximately 3 years in our previous work [8]. The design of this time-varying voiceprint database cleared out speaker-independent mismatches and mismatches in speaking style and speech contents by relevant control measures. Preliminary experimental results showed that speaker recognition system performed best when training and testing utterances are from the same session, i.e., on the same

recording dates. However, the performance gets worse and worse with the recording date difference between training and testing gets bigger [8].

This result serves as a possible proof for the efficiency of the model updating method. The shortcoming of this method is also evident, as it is costly, user-unfriendly and sometimes may be unrealistic for real applications.

Performance improvement against mismatches always resorts to selecting better features. In time-varying speaker recognition, the most essential way to stabilize performance is to extract exact acoustic features that are speaker-specific and further, stable across sessions. Acoustic parameters, such as pitch, formant, have been examined first, while it seems they remain more or less the same across sessions and no valuable trend has been tracked so far. Efforts have also been made in the frequency domain, where we have been trying to identify frequency bands that reveal high discrimination sensitivity for speaker-specific information but low discrimination sensitivity for session-specific information. Once these frequency bands are identified, more features can be extracted within them by means of frequency warping. Thus information critical to time-varying speaker recognition is emphasized and performance improvement can be expected. A discrimination score for each frequency band can be obtained regarding the requirements analyzed above, and frequency warping is done on the basis of the classic Mel scale, which is named as a Discrimination-emphasized Mel-frequency-warping method.

This paper is organized as follows. In Section II, the proposed method for time-varying speaker recognition is detailed. A brief description of the time-varying voiceprint database is presented in Section III. Experimental setup and results are listed in Section IV. Conclusions are drawn in Section V.

II. The Discrimination-emphasized Mel-frequency-warping Method

As analyzed in Section I, the proposed solution is to highlight in feature extraction the frequency bands that reveal high discrimination sensitivity for speaker-specific information while low discrimination sensitivity for session-specific information. Then this problem split into two sub-problems: how to determine the discrimination sensitivity of each frequency band in this time-varying speaker recognition task and how to do frequency warping to highlight target frequency bands.

A. Discrimination Score Calculation

Out of those discriminate criteria in machine learning, F-ratio has broadly served as a criterion of feature selection in speaker recognition [9], which is the ratio of the between-group variance to the within-group variance. A higher F-ratio value means better feature selection for the target grouping. That is to say, the feature selection with a higher F-ratio possesses higher discrimination sensitivity against the target grouping [10].

This idea is employed to determine the importance of frequency bands in time-varying speaker recognition. The whole frequency range is divided into *K* frequency bands uniformly and linear frequency scale triangle filters are used to process the power spectrum of utterances. The filter-setup is the same as that of classic MFCC (Mel-frequency Cepstrum Coefficients) except for linear frequency scaling.

Suppose there are *M* speakers and *S* sessions in a given database. In this case, there are two different kinds of grouping: grouping by speakers for each session and grouping by sessions for each speaker, which correspond to two different kinds of F-ratios.

The first kind of F-ratio, denoted as *F-ratio-spk*, is illustrated in Equ. (1):

$$F\text{-}ratio\text{-}spk_s^k = \frac{\sum_{i=1}^{M}\left(\mu_{i,s}-\mu_s\right)^2}{\sum_{i=1}^{M}\frac{1}{N_{i,s}}\sum_{j=1}^{N_{i,s}}\left(x_{i,s}^{k,j}-\mu_{i,s}\right)^2}, \tag{1}$$

where $F\text{-}ratio\text{-}spk_s^k$ denotes the F-ratio value of frequency band *k* in session *s*, $x_{i,s}^{k,j}$ is power of the frequency band *k* in frame *j* of the speaker *i* in session *s*, $N_{i,s}$ is the frame number of speaker *i* in session *s*, and $\mu_{i,s}$ and μ_s are corresponding averages calculated as follows.

$$\mu_{i,s} = \frac{1}{N_{i,s}}\sum_{j=1}^{N_{i,s}} x_{i,s}^{k,j}. \tag{2}$$

$$\mu_s = \frac{1}{M}\sum_{i=1}^{M}\mu_{i,s}. \tag{3}$$

For each frequency band *k*, there is an averaged $F\text{-}ratio\text{-}spk^k$:

$$F\text{-}ratio\text{-}spk^k = \frac{1}{S}\sum_{s=1}^{S} F\text{-}ratio\text{-}spk_s^k. \tag{4}$$

Frequency bands with higher *F-ratio-spk* have higher discrimination sensitivity for speaker-specific information.

Similarly, the second kind of F-ratio, denoted as *F-ratio-ssn* , is illustrated in Equ. (5):

$$F\text{-}ratio\text{-}ssn_i^k = \frac{\sum_{s=1}^{S}\left(\mu_{i,s}-\mu_i\right)^2}{\sum_{s=1}^{S}\frac{1}{N_{i,s}}\sum_{j=1}^{N_{i,s}}\left(x_{i,s}^{k,j}-\mu_{i,s}\right)^2}. \tag{5}$$

where $F\text{-}ratio\text{-}ssn_i^k$ denotes the F-ratio value of frequency band *k* of speaker *i*, and μ_i is the average calculated as follows.

$$\mu_i = \frac{1}{S}\sum_{s=1}^{S}\mu_{i,s}. \tag{6}$$

For each frequency band *k*, there is an averaged $F\text{-}ratio\text{-}ssn^k$:

$$F\text{-}ratio\text{-}ssn^k = \frac{1}{M}\sum_{i=1}^{M} F\text{-}ratio\text{-}ssn_i^k. \tag{7}$$

Frequency bands with lower *F-ratio-ssn* have lower discrimination sensitivity for session-specific information.

Then for each frequency band *k*, a discrimination score $discrim\text{-}score^k$ can be defined as:

$$discrim\text{-}score^k = \frac{F\text{-}ratio\text{-}spk^k}{F\text{-}ratio\text{-}ssn^k}. \quad (8)$$

B. Mel-frequency-warping Strategies

Mel scale, one kind of frequency warping, takes into account human auditory characteristics and has been the state-of-the-art technology in feature extraction in both speech and speaker recognition, which is the basis of our proposed warping method.

One warping strategy is to uniformly warp those target frequency bands with discrimination scores above a threshold. Warping-factors are designed to emphasize information within target frequency bands, as they contribute more to the time-varying speaker recognition task. Evidently this does not mean non-target frequency bands are of no contribution to the task. Therefore, target frequency bands should be assigned with a proper warping-factor, neither too small to add emphasis, nor too big to pose a threat to the whole system. The proposed frequency warping, called Mel-frequency-warping (MFW), is illustrated in Fig. 1.

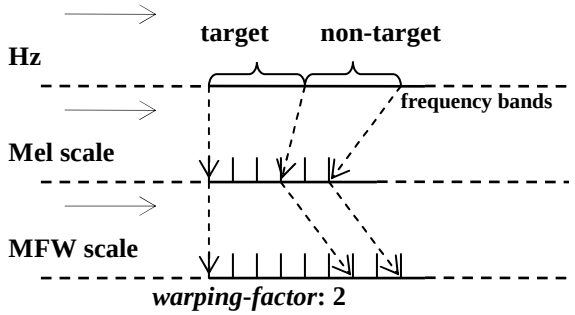

Fig. 1.The relationship between Hz, Mel scale and MFW scale

In a time-varying speaker recognition system, there may be several discontinuous target frequency bands with their discrimination scores higher than a specified threshold. In this case, the warped Mel frequency becomes complicated while the warping principle remains the same: processing Mel frequency in an increasing order with target frequency bands warping by a certain factor (>1) and non-target frequency bands unchanged (warping factor is 1).

A comparison of the extraction procedures of MFCC and the proposed WMFCC (Warped Mel-frequency Cepstrum Coefficients) is shown in Fig. 2. Clearly, MFCC is the same as WMFCC with warping factors of all frequency bands being 1.

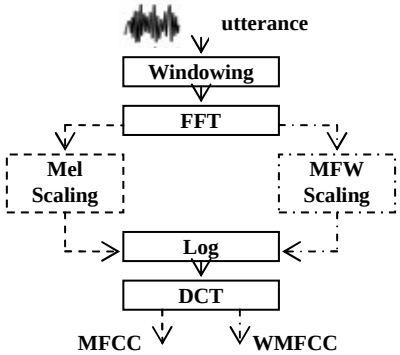

Fig. 2. A comparison of MFCC and WMFCC extraction procedures

Since the warping-factor represents the degree of information emphasis by Mel-frequency-warping, the value of a discrimination score can be a reference in choosing its corresponding warping-factor. Thus another warping strategy is non-uniformly warping of the whole frequency range according to their discrimination scores. This strategy requires a more complicated determination procedure of warping-factors, which is to be done in the future.

III. The Time-varying Voiceprint Database

The time-varying voiceprint database [8] is used in the research, which aims to contribute to examining solely the time-varying impact on speaker recognition. To avoid mismatches other than time-related variability, recording equipments (microphone-channel), software, conditions and environment are kept as constant as possible. Furthermore, speakers are requested to utter in a reading way with fixed prompt texts (100 Chinese sentences with varied lengths) instead of free-style conversations (employed in the MARP corpus [11]) throughout 16 sessions in a period of approximately three years (from 2010 to 2012). Sessions are of gradient time intervals where initial ones are of shorter time intervals and following ones of longer and longer time intervals. All speakers are recruited on campus, with 30 female and 30 male.

Following experiments are performed on 8kHz-sampling microphone data from the first 10 recording sessions. The 10th session was recorded approximately a year away from the 1st one.

IV. Experimental Results

A. Experimental Setup

All experiments were based on the state-of-the-art 1024-mixture GMM-UBM (Gaussian Mixture Model – Universal Background Model) speaker recognition system. 16-dimensional MFCCs and their first derivatives were adopted as acoustic features in the baseline system, while 16-dimensional WMFCCs and their first derivatives in the proposed method.

Each speaker model was trained using 3 sentences randomly selected from the entire 100 sentences from the 2nd session with length of about 10 seconds, and all sentences from the first 10 sessions were used for testing with each sentence ranging from 2 to 5 seconds.

B. Determination of WMFCC Parameters

The whole frequency-range (from 100 Hz to 3800 Hz) was divided into 30 frequency bands. All data from the first 10 sessions were used to calculate the discrimination score of each frequency band, as shown in Fig. 3.

As can be seen from the figure, below 2500 Hz, the discrimination scores generally fluctuated within the range of 2 to 3, while above 2500 Hz, the curve climbed up with all values well above 3, which was the average. Hence, 2500 Hz ~ 3800 Hz was identified as the target frequency bands. A

series of experiments had been done to find a proper warping-factor and it came out that the system performed best with a warping-factor of 3 as shown in Table I.

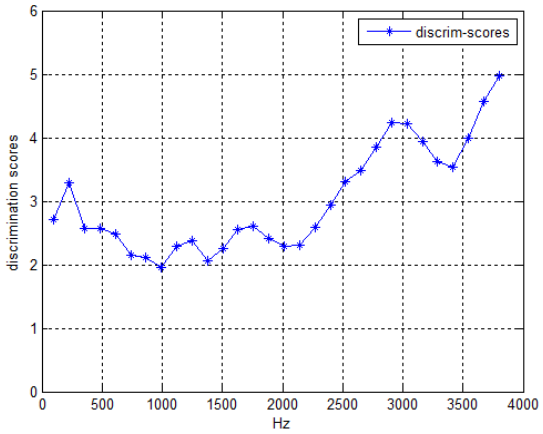

Fig. 3. Discrimination scores of frequency bands

TABLE I
A COMPARISON OF THE PERFORMANCE OF WMFCC WITH DIFFERENT WARPING FACTORS IN AVERAGE EER (%)

warping factor	1	2	3	4	5
WMFCC	10.06	8.69	8.14	8.22	8.36

C. *Experimental Results*

Choose 2500Hz~3800Hz as target frequency bands and 3 as the warping factor. Fig. 4 shows a comparison of the performance of MFCC and proposed WMFCC in EER (%), while Table II presents another comparison in degradation degree with time varying in average (%) and the reduction rate (RR, %).

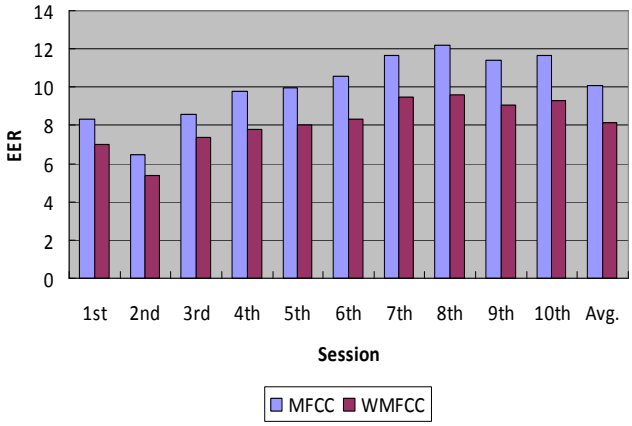

Fig. 4. A comparison of the performance of MFCC and WMFCC in EER

TABLE II
ANOTHER COMPARISON OF THE PERFORMANCE OF MFCC AND WMFCC IN DEGRADATION DEGREE WITH TIME VARYING

	2nd-session EER	Average EER	Average Degradation Degree
MFCC	6.45	10.06	0.56
WMFCC	5.38	8.14	0.51
RR	16.6	19.1	8.9

Fig. 4 clearly demonstrates the time-varying effect on speaker recognition with the 2nd session performed the best where training utterances were selected. After about half a year, EERs generally fluctuated around 12%. Since the proposed feature of WMFCC took into account both speaker-specific information and session-specific information, it yielded a reduction of 19.1% in average EER, and also a reduction of 8.9% in average degradation degree with time varying.

V. CONCLUSIONS

A Discrimination-emphasized Mel-frequency-warping method is proposed in this paper for time-varying speaker recognition. Experimental results show that in the time-varying voiceprint database, this method can not only improve speaker recognition performance in average EER with a reduction of 19.1%, but also alleviate performance degradation brought by time varying with a reduction of 8.9%.

Further experiments are needed to test the data-dependency by using other databases.

Also, it requires more speculation and experimentation whether the discrimination-emphasized idea could be applied to other speech features, and further, speaker modeling techniques.

REFERENCES

[1] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, iss. 9, pp. 859-872, September 1997.

[2] H. J. Kunzel, "Current approaches to forensic speaker recognition," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 135-141, 1994.

[3] L. G. Kersta, "Voiceprint recognition," *Nature*, no. 4861, pp. 1253-1257, December 1962.

[4] J. Bonastre, F. Bimbot, L. Boe, *et al*., "Person authentication by voice: a need for caution," *Proc. of Eurospeech 2003*, pp. 33-36, Geneva, 2003.

[5] F. Soong, A. E. Rosenberg, L. R. Rabiner, *et al*., "A vector quantization approach to speaker recognition," *Proc. of ICASSP 1985*, vol. 10, pp. 387-390, Florida, 1985.

[6] T. Kato, and T. Shimizu, "Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence preserving connected digit patterns," *Proc. of ICASSP 2003*, Hong Kong, 2003.

[7] M. Hebert, "Text-dependent speaker recognition," *Springer Handbook of Speech Processing*, Springer-Verlag: Berlin, 2008.

[8] L. Wang and T. F. Zheng, "Creation of time-varying voiceprint database," *Proc. of O-COCOSDA 2010*, Kathmandu, 2010.

[9] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Am*, vol. 51, no. 6, pp. 2044-2056, 1972.

[10] X. Lu and J. Dang, "Physiological feature extraction for text independent speaker identification using non-uniform subband processing," *Proc. of ICASSP 2007*, pp. 461-464, 2007.

[11] A. D. Lawson, A. R. Stauffer, E. J. Cupples, *et al*., "The multi-session audio research project (MARP) corpus: goals, design and initial findings," *Proc. of Interspeech 2009*, pp. 1811-1814, Brighton, 2009.